Enfoque UTE, V.11-N.1, Ene.2020, pp. 56-67
http://ingenieria.ute.edu.ec/enfoqueute/
e-ISSN: 1390-6542 / p-ISSN: 1390-9363

Recibido (Received): 2019/11/22
Aceptado (Accepted): 2020/01/17

Detección de valores atípicos con técnicas de minería de datos y métodos estadísticos

(Outlier detection with data mining techniques and statistical methods)

Marcos Orellana,[1] *Priscila Cedillo*[2]

Resumen

La detección de valores atípicos en el campo de la minería de datos (DM) y el descubrimiento de conocimiento a partir de datos (KDD) es de gran interés en áreas que requieren sistemas de soporte a la toma de decisiones, como, por ejemplo, en el área financiera, en donde mediante DM se pueden detectar fraudes financieros o encontrar errores producidos por los usuarios. Entonces, es esencial, evaluar la veracidad de la información, a través de métodos de detección de comportamientos inusuales en los datos. Este artículo propone un método para detectar valores que se consideran valores atípicos en una base de datos de datos de tipo nominal. El método implementa un algoritmo global de "k" vecinos más cercanos, un algoritmo de agrupamiento denominado *k-means* y un método estadístico denominado chi-cuadrado. La aplicación de estas técnicas ha sido implementada sobre una base de datos de clientes que han solicitado un crédito financiero. El experimento se realizó sobre un conjunto de datos con 1180 tuplas, en donde, deliberadamente se introdujeron valores atípicos. Los resultados demostraron que el método propuesto es capaz de detectar todos los valores atípicos introducidos.

Palabras clave

Valor atípico; minería de datos; KNN; chi-cuadrado; fraude financiero.

Abstract

The detection of outliers in the field of data mining (DM) and the process of knowledge discovery in databases (KDD) is of great interest in areas that require support systems for decision making. A straightforward application can be found in the financial area, where DM can potentially detect financial fraud or find errors produced by the users. Thus, it is essential to evaluate the veracity of the information, through the use of methods for the detection of unusual behaviors in the data. This paper proposes a method to detect values that are considered outliers in a database of nominal type data. The method implements a global algorithm of "k" closest neighbors, a clustering algorithm called k-means and a statistical method called chi-square. These techniques have been implemented on a database of clients who have requested a financial credit. The experiment was performed on a data set with 1180 tuples, where, outliers were deliberately introduced. The results showed that the proposed method is able to detect all the outliers entered.

Keywords

Outlier, data mining, KNN, chi-square, financial fraud.

1. Introducción

La detección de valores atípicos representa un desafío en las técnicas de minería de datos. Los valores atípicos o también denominados anómalos tienen propiedades diferentes con respecto a la generalidad, ya que debido a la naturaleza de sus valores y por ende, a su comportamiento no son datos que mantienen un comportamiento similar a la mayoría. Los datos anómalos son susceptibles de ser introducidos por mecanismos maliciosos (Atkinson, 1981). Mandhare y Idate (2017) consideran que este tipo de datos son una amenaza y los definen como irrelevantes o

1 Universidad del Azuay, Cuenca-Ecuador (marore@uazuay.edu.ec).
2 Universidad de Cuenca, Cuenca-Ecuador (priscila.cedillo@ucuenca.edu.ec).

maliciosos. Además, estos datos generan conflictos durante el proceso de análisis, lo que resulta en información poco confiable e inconsistente. Sin embargo, si bien los datos anómalos son irrelevantes para encontrar patrones en la cotidianidad de los datos, son útiles como objeto de estudio en casos en donde, mediante estos es posible identificar mediante un proceso no controlado, problemas tales como fraudes financieros.

El proceso de detección usando técnicas de minería de datos, facilita la búsqueda de valores anómalos o patrones inusuales en los datos (Arce, Lima, Orellana, Ortega y Sellers, 2018). Varios estudios muestran que la mayoría de este tipo de datos se originan también en dominios como tarjetas de crédito (Bansal, Gaur y Singh, 2016), sistemas de seguridad (Khan, Pradhan y Fatima, 2017) e información de salud electrónica (Zhang y Wang, 2018).

El proceso de detección incluye un proceso de minería de datos que utiliza herramientas basadas en algoritmos de tipo no supervisado (Onan, 2017). El proceso de detección consta de dos enfoques según su forma: local y global (Monamo, Marivate y Twala, 2017). Los enfoques globales incluyen un conjunto de técnicas en las que se asigna una puntuación a cada anomalía con relación al conjunto de datos globales. Por otro lado, los enfoques locales, representan las anomalías en un dato determinado con respecto a su vecindad directa; es decir, a los datos cercanos en cuanto a la similaridad de sus características. De acuerdo a los conceptos antes mencionados, el enfoque local detecta valores atípicos que son ignorados cuando se utiliza un enfoque global, en especial en aquellos con densidad variable (Amer y Goldstein, 2012). Ejemplos de dichos algoritmos son aquellos basados en i) agrupamiento y ii) el vecino más cercano. El algoritmo de la primera categoría considera que los valores atípicos están en vecindarios dispersos, que están lejos de los vecinos más cercanos. Mientras que la segunda categoría, opera en los algoritmos agrupados (Onan, 2017).

En otros trabajos se analizan las anomalías considerando varias columnas de la tupla, mas no la detección a nivel de cada columna (Dang, Ngan y Liu, 2015; Ganji, 2012; Gu et al., 2017; Malini y Pushpa, 2017; Mandhare y Idate, 2017; Sumaiya Thaseen y Aswani Kumar, 2017; Yan, You, Ji, Yin y Yang, 2016). Kuna et al. (2013) presenta un procedimiento que utiliza árboles de decisión basados en el algoritmo C4.5, el cual aplica valores continuos y separa los posibles resultados en dos ramas. El método genera un árbol de decisión con estos datos, por medio de particiones obtenidas recursivamente. Para encontrar los atributos más significativos y luego cada atributo de entrada y salida, se aplica el factor de valor atípico mediante la técnica de Local Outlier Factor (*LOF*). Si la metodología es efectiva, el nuevo algoritmo debe ser evaluado, basado el concepto de "vecinos más cercanos" para mejorar la precisión de la solución. Existen otros estudios que demuestran hasta un 75 % de mejora en el rendimiento de la clasificación de este algoritmo (Rosero-Montalvo et al., 2018). No solamente existe un método; además, también se han presentado estudios que proveen métodos híbridos (p. ej., métodos adaptativos y basados en grupos) que aceleran el algoritmo de clasificación. En estos últimos, se considera especialmente el tiempo de respuesta para procesos sobre grandes conjuntos de datos (Ougiaroglou, Evangelidis y Dervos, 2014).

Consecuentemente, el presente artículo, propone una metodología para la detección de valores atípicos, que se basa en la aplicación de métodos estadísticos tradicionales (prueba de chi-cuadrado) con la aplicación de algoritmos de minería de datos (*KNN Global Anomaly* y K-means). La experimentación del método se ha aplicado a un dominio financiero. Finalmente, se realizó una evaluación en la que se obtuvieron buenos resultados sobre la aplicación de la metodología propuesta en este artículo y su desempeño en el tratamiento de valores atípicos, incluso cuando hay valores nulos.

Este estudio está estructurado de la siguiente manera: la Sección 2 discute las soluciones existentes. La Sección 3 presenta el método propuesto. La Sección 4 presenta la evaluación de la aplicación de la metodología, y finalmente, la Sección 5 presenta las conclusiones y el trabajo futuro.

2. Trabajos relacionados

Existen varios enfoques relacionados con la detección de valores atípicos, en este contexto, Hassanat, Abbadi, Altarawneh, y Alhasanat (2015), realizaron una encuesta en donde se presenta un resumen de los diferentes estudios de detección de valores atípicos, siendo estos: el enfoque basado en estadísticas, el enfoque basado en la distancia y el enfoque basado en la densidad. Los autores presentan una discusión relacionada con los valores atípicos, los métodos utilizados para agrupar el conjunto de datos y, concluyen con que el algoritmo k-mean es el más popular en la agrupación de un conjunto de datos. Además, en otros estudios (Dang et al., 2015; Ganji, 2012; Gu et al., 2017; Malini y Pushpa, 2017; Mandhare y Idate, 2017; Sumaiya Thaseen y Aswani Kumar, 2017; Yan et al., 2016) se utilizan técnicas de minería de datos, métodos estadísticos o ambos. Para la detección de valores atípicos, comúnmente se han aplicado las técnicas del vecino más cercano (KNN) junto con otras para encontrar patrones inusuales durante el comportamiento de los datos o para mejorar el rendimiento del proceso. Gu et al. (2017) presentan un método eficiente basado en la cuadrícula para encontrar patrones de datos atípicos en grandes conjuntos de datos. Del mismo modo, Yan et al. (2016) proponen un método de detección atípico con KNN y poda de datos, el cual toma muestras sucesivas de tuplas y columnas, y aplica un algoritmo KNN para reducir la dimensionalidad sin perder información relevante. Con relación al uso de técnicas de minería de datos aplicadas al sector financiero, el estudio presentado por Malini y Pushpa (2017) explica la detección atípica de fraudes de tarjetas de crédito con técnicas que detectan patrones inusuales a través del algoritmo KNN. La técnica detecta la los datos de entrada, calcula el vecino más cercano y el puntaje de similitud, determinando si los datos tienen una sospecha de fraude. Ganji (2012) presenta un estudio similar que aplica KNN mediante el uso de una técnica que utiliza una ventana con memoria asignada, de esta manera, cuando se detecta una transacción de solicitud, solo necesita un escaneo de la ventana actual, para encontrar el objeto cuyos k vecinos más cercanos están influenciados.

Estudios relacionados también utilizan el estadístico chi-cuadrado, tal es el caso de Sumaiya Thaseen y Aswani Kumar (2017), cuyo método implementa el estadístico para la detección de intrusos en una red informática para segmentarlo en tráfico normal y anormal. El estudio incluye un método que utiliza la técnica *Support Vector Machine* (SMV).

No se encontraron estudios que combinen el método estadístico chi-cuadrado con KNN. Es decir, un método que pueda ayudar a clasificar previamente las columnas relevantes antes de aplicar la técnica KNN. Esta propuesta plantea una metodología híbrida que primero selecciona las columnas más relevantes en cuanto a la relación columna entrada-columna salida y luego aplica KNN. Sin embargo, el estudio propuesto, a diferencia de los mencionados, utiliza una variación KNN llamada *KNN Global Anomaly*, que tiene adicionalmente posee una medida de anomalía

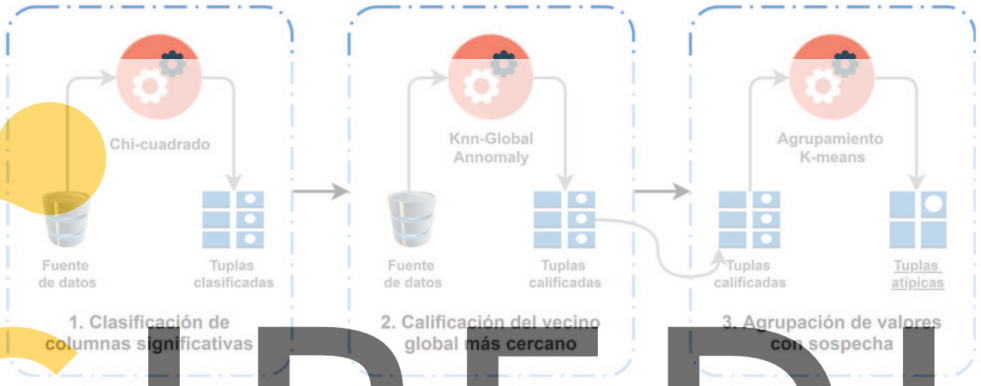
3. Metodología

En esta sección, se presenta la metodología para encontrar valores atípicos en una base de datos nominal (conjunto de datos). Esta propuesta evaluó tuplas y columnas (perfil del cliente) con respecto a la columna de salida (monto de crédito otorgado) para encontrar valores atípicos. La

columna de salida denominada *S* es un valor discretizado que representa los valores obtenidos en función de un conjunto de columnas de entrada.

El método propuesto se dividió en tres fases principales que se ilustran en la figura 1: i) Uso del método estadístico chi-cuadrado para definir una ponderación o peso *W* que indica la relevancia de la columna del perfil del cliente en comparación con la cantidad de crédito otorgado, ii) en la siguiente fase, se utilizó el algoritmo de agrupación denominado *KNN Global Anomaly* (KNN) que calcula la puntuación de los valores atípicos, y iii) finalmente, el algoritmo K-Means que es utilizado para separar los valores atípicos de los valores falsos-positivos.

Figura 1. Metodología para la detección de valores atípicos



En las subsecciones siguientes, se describen cada una de las actividades:

Clasificación de columnas significativas

Para clasificar las columnas significativas, se utilizó el estadístico chi-cuadrado. Chi-cuadrado es una prueba no paramétrica utilizada para determinar si una distribución de frecuencias observadas difiere de las frecuencias teóricas esperadas (Gol y Abur, 2015). El peso de la columna de entrada (columnas que determinan el perfil del cliente) se calcula con relación a la columna de salida (monto del crédito). Cuanto mayor es el peso de una columna correspondiente a las columnas de entrada en una escala de cero a uno, más relevante se considera. Es decir, mientras el valor del peso se acerque más a uno, la relación con respecto a la columna de salida será más importante. El estadístico puede ser aplicado solamente a columnas de tipo nominal y ha sido seleccionado como método para definir relevancias. Chi-cuadrado reporta un nivel de significancia de las asociaciones o dependencias y se utilizó como prueba de hipótesis sobre el peso o importancia de cada una de las columnas con respecto a la columna de salida *S*. El valor resultante se almacena en una columna denominada *weight*, que junto al puntaje de anomalía se reportan al final del proceso.

Calificación del vecino local más cercano

Para obtener los valores con sospecha de anormalidad, se utiliza K-NN Global Anomaly Score. KNN se basa en el algoritmo del k vecino más cercano, el cual calcula el puntaje de anomalía de los datos en relación con el vecindario. Usualmente, los valores atípicos están lejos de sus vecinos o su vecindario es escaso. En el primer caso, se conoce como detección de anomalías

global y se lo identifica con KNN; el segundo, se refiere a un enfoque basado en densidad local. El puntaje proviene por defecto de la media de la distancia a los vecinos más cercanos (Amer y Goldstein, 2012). En la clasificación del *k* vecino más cercano, la columna de salida *S* del vecino más cercano del dataset de entrenamiento, se relaciona con un nuevo dato no clasificado en la predicción, esto implica una línea de decisión lineal.

Para obtener una predicción correcta, el valor *k* (número de vecinos a considerarse alrededor del valor analizado), debe ser configurado cuidadosamente. Un valor alto de *k* representa una mala solución con respecto a la predicción, mientras valores bajos tienden a generar ruido (Bhattacharyya, Jha, Tharakunnel y Westland, 2011). Frecuentemente, el parámetro *k* es elegido empíricamente y depende de cada problema. Hassanat, Abbadi, Altarawneh, y Alhasanat (2014) proponen la realización de pruebas con diferentes números de vecinos cercanos hasta llegar al que posee mejor precisión. Su propuesta inicia con valores desde k=1 hasta k= raíz cuadrada del número de tuplas del dataset de entrenamiento. La regla general, a menudo es asignar *k* con la raíz cuadrada del número de tuplas del dataset *D*.

Tomando en cuenta las consideraciones anteriores, se configuró k=34 para el dataset; el valor resultante utilizó la ecuación 1, donde *n* identifica al número de tuplas del dataset y TRUNC a la función de truncamiento del valor obtenido de la raíz cuadrada.

$$k = TRUNC\ (\sqrt{n}) \tag{1}$$

El proceso KNN incluye un procedimiento de preparación de datos, el cual convierte la columna de análisis en una tupla. Es decir, el nombre de la columna forma parte de las tuplas en una columna nueva denominada *attribute*. La ejecución del algoritmo KNN, genera una columna denominada *outlier*; la cual registra el peso de la anomalía de los datos en relación con todas las tuplas del conjunto de datos *D*. El resultado de la columna *outlier* varía de acuerdo a la distancia de cada uno de los datos, con respecto a la distancia media de los vecinos más cercanos. Las tuplas que poseen valores superiores a cero son aquellas que tienen probabilidades de clasificarse como valores anómalos, no obstante, debido a la variabilidad de valores obtenidos en cada iteración y, para mantener uniformidad en el criterio de selección para las columnas examinadas, los valores fueron normalizados a una escala entre cero y uno. El valor cero corresponde a las tuplas que no tienen probabilidad de contener valores anómalos. Si el valor de *outlier* es mayor a cero y tiende cada vez más a uno, corresponde a que la tupla tiene mayores probabilidades de ser un valor anómalo. El valor de la columna *outlier* fue incluido en reporte final para un análisis detallado.

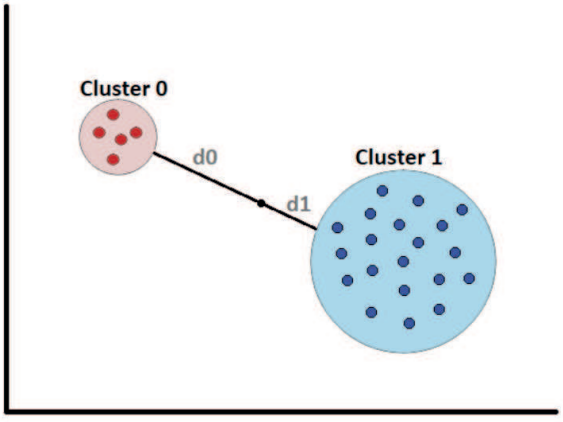
La precisión de un algoritmo KNN no siempre resulta óptima en el análisis de todas las columnas. La cardinalidad es una característica considerada como relevante; además, cada columna varía en relación directa con el género de información. Es decir, en el caso de valores discretos con baja cardinalidad, la precisión aumenta, debido a que la frecuencia se concentra en pocos valores. Como resultado, valores no clasificados en dichas concentraciones, fueron considerados como valores anómalos. Por ejemplo, la columna sexo tiene una cardinalidad de dos (“M” Masculino, “F” Femenino). Los valores no coincidentes con estos dos grupos, se clasificaron por K-NN con un puntaje de uno en la columna *outlier*, y en consecuencia calificados como valores anómalos. El siguiente caso son los valores con alta cardinalidad, donde se forman muchas clases con frecuencias menores. En este caso, el algoritmo K-NN disminuye su precisión en la separación de clases (normales, anómalos) debido a la alta dispersión. Por ejemplo: la columna edad tiene una alta cardinalidad por el alto número de clases que se crean.

Agrupación de valores con sospecha

Finalmente, para agrupar o separar los valores atípicos se utiliza el algoritmo K-means. Este algoritmo se basa en distancias y es utilizado para dividir los datos en un número de clústeres (Aldahdooh y Ashour, 2013). Los datos agrupados en un mismo clúster son similares comparados con otros clústeres. K-means asigna aleatoriamente un punto denominado centroide basado en el valor del parámetro *k*, desde el cual se toman distancias a los otros puntos. La métrica de distancia puede ser: euclidiana, coseno o distancia coseno rápida (para este caso se utilizó distancia euclidiana). Una vez calculada la distancia, se asigna iterativamente cada dato al clúster correspondiente de acuerdo a la similitud en función de la distancia y el valor medio de los datos. A continuación, se vuelve a calcular los centroides y el valor medio de la distancia a los datos hasta que los centroides no se muevan (Sinwar y Dhaka, 2015).

El algoritmo de K-means garantiza el resultado del método anterior (K-NN) en especial las columnas de alta dispersión. Para el proceso se aplica una validación por distancia, esta consiste en dividir el conjunto de datos de cada columna en dos clústeres. Un primer clúster para falsos-positivos y otro de valores atípicos. Para lograr esto se ejecuta el algoritmo k-means con el parámetro k=2, el cuál indica que se formarán dos grupos. Un proceso de medición de rendimiento fue requerido, el cual produce un vector de distancia: distancia media global, distancia media al clúster cero y distancia media al clúster uno. Entre el punto que marca la distancia media de los clústeres y cada clúster, existe una distancia que separa los valores atípicos de los falsos-positivos. El clúster con mayor distancia a la media global corresponde al clúster de valores atípicos. Los cálculos fueron realizados con un índice denominado *distance_factor*, el cuál puntúa entre rangos de cero y uno. Estos valores corresponden a la proporcionalidad de las distancias con el grado de alejamiento entre los clústeres y la media global. Es decir, una relación significativa entre la distancia media general y la distancia del clúster cero, en comparación con la distancia promedio general y la distancia del clúster uno; esto determina el clúster que posee los valores atípicos de aquellos que son falsos positivos (ver figura 2).

Figura 2. Distancia entre clústeres para separación de falsos positivos



Al inicio del proceso, se asignó un puntaje a cada par de columnas. De este par, una de las columnas es denominada columna de entrada y corresponde al perfil del cliente; la otra es denominada columna de salida u objetivo que corresponde al monto de crédito. Los valores almacenados en la columna *weigth* son producto del cálculo del estadístico chi-cuadrado en un rango

de cero a uno y establece la relevancia de cada columna con respecto a la salida *S*. Valores con tendencia a uno son aquellos más relevantes en cuanto a su influencia sobre S. Una vez que se ejecuta K-NN y K-means, se introduce un filtro que limita el reporte de salida basado en el peso de la anomalía. Inicialmente se colocan solo aquellos valores que superen un peso mayor a cero.

Para evaluar el método de detección de valores atípicos, se generó un conjunto de datos en el que se introdujeron valores que representan anomalías de acuerdo con los criterios de clasificación y frecuencia. Para incluir esos valores, se han realizado los siguientes pasos.

1. Identificar un número suficiente de características para construir los criterios de clasificación.
2. Determinar según la frecuencia, la métrica para establecer la división entre los datos.
3. Establecer un criterio basado en los resultados de las frecuencias de las observaciones, que se clasifican como valores atípicos.

La inclusión de valores atípicos se completó de acuerdo con el enfoque mencionado anteriormente, identificando las tuplas y los valores de las columnas con sospecha de anomalía. El objetivo fue la comparación de los elementos de la fase de identificación con los resultados del procedimiento.

Las tuplas del conjunto de datos corresponden a una base de datos de créditos de tipo nominal con 15 columnas como se indica en la tabla 1. Un total de 13 columnas corresponden a características personales del solicitante (perfil del cliente); otra columna, al identificador de la cuenta "account_id", y finalmente "credit_granted" representa el valor discretizado del monto de crédito. Esta última, también se denomina: columna salida u objetivo. El conjunto de datos proviene de una demostración utilizada en ambientes de pruebas para encontrar valores anómalos; tiene 1180 tuplas y contiene valores nulos.

Tabla 1. Columnas del set de datos

Variable	Concepto
account_id	Identificador de la cuenta
age	Edad
gender	Genero
antiquity	Antigüedad
is_propietary	Es propietario
civil status	Estado civil
children	Tiene hijos
credit Card	Tarjeta de crédito
credit_card_qty	Cantidad de tarjetas de crédito
country	País
anual_income	Ingresos anuales
education	Nivel de educación
job	Ocupación
credit granted	Crédito otorgado
status	Estado

El *software* para la experimentación fue *Rapid Miner Studio* en la versión 8.1.01. Adicionalmente, se instaló la extensión *Anomaly Detection versión 2.4.001 de German Research for Artificial Intelligence*.

El conjunto de datos fue verificado con respecto a coherencia y veracidad. Se observó que el número de cuenta que representa el denominado "account_id" contiene valores repetidos y alterados; para ello, se generó un nuevo identificador alterno. Durante el procedimiento de experimentación se analizó de forma detallada cada una de las columnas de la base de datos con respecto a la información que esta contiene, para ello se implementaron tablas de frecuencia. Los valores con baja frecuencia con relación al resultado global tuvieron la principal sospecha de pertenecer al grupo de valores anómalos. La observación también examinó valores relacionados al patrón de comportamiento del contexto; se determinó los candidatos a valores anómalos. Los valores obtenidos se indagaron en cada experimentación para verificar la efectividad del procedimiento.

Validación

Con el fin de obtener un resultado confiable se realizó un proceso de validación en el cual, se ejecutó previamente un levantamiento de la base de datos columna a columna, agrupando cada uno de los valores, sin incluir valores repetidos. A cada uno de los valores le acompaña la frecuencia de repetición que establece la primera indagación de anomalía. Por ejemplo, si se analiza la columna "gender"; 605 de los valores son "F", 574 son "M", y un valor contiene el valor "4", se podría asegurar que el valor anómalo en el procedimiento debería identificar es el valor "4" y no a "M" que corresponde seguramente a masculino y "F" a femenino.

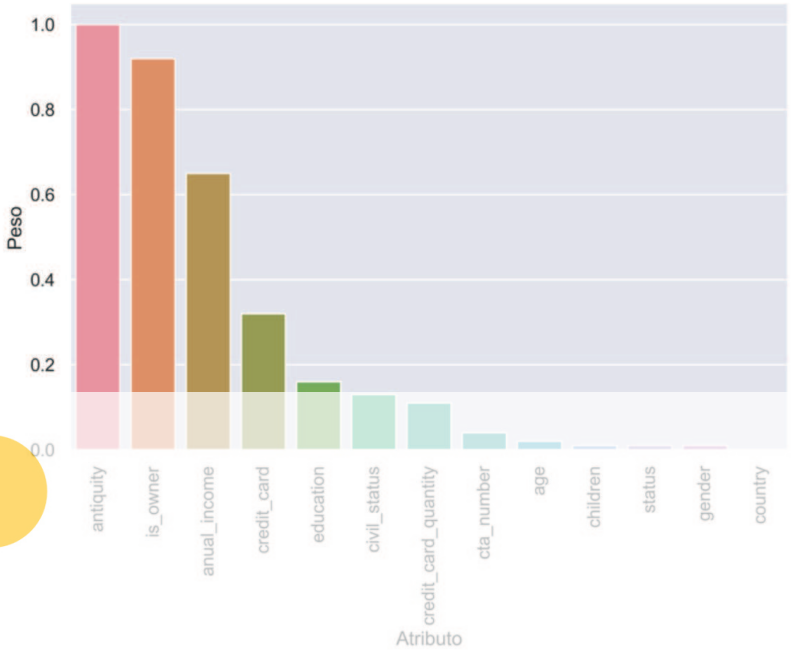
4. Resultados y discusión

En el intento detectó eficientemente cierto dellos valores anómalos y falsos positivos de una base de datos para concesión de créditos financieros.

La ejecución del método estadístico chi-cuadrado permitió definir los atributos que no aportan un grado de significancia de decisión con relación a la columna de salida denominada "credit_granted". La columna "country" no define relevancia, a diferencia de las columnas: "antiquity", "is_propietary", "anual_income" y "job", con valores de peso superiores a 0.30 lo que determina mayor significancia con respecto a las columnas con menos valoración. Las columnas restantes con valores mayores a cero, pese a que no tienen una relevancia superior, se consideran dentro del procedimiento para la búsqueda de valores anómalos. La figura 3 ilustra los valores resultantes de la aplicación de Chi-cuadrado.

El análisis de valores atípicos para la columna "account_id" no fueron considerados en el estudio, ya que presentan inconsistencias en términos de codificación. Cinco tuplas tienen asignadas un valor de menos uno. El atributo tampoco es considerado puesto que tiene la calidad de identificador primario, y por ende contiene alta variabilidad, lo cual no es importante para determinar la columna de salida.

Figura 3. Valores Resultantes de la Aplicación de Chi-cuadrado



La tabla 3 muestra los resultados obtenidos en la aplicación del procedimiento con respecto al estudio inicial de las tablas de frecuencias.

Tabla 3. Columnas del set de datos.

Atributos	Anómalos original	Anómalos detectados	Datos nulos	Efectividad
antiquity	1	3	0	100 %
is_owner	1	1	0	100 %
anual_income	0	6	0	100 %
job	1	1	0	100 %
credit_card	0	1	0	100 %
education	2	2	0	100 %
civil_status	2	2	0	100 %
credit_card_quantity	2	2	0	100 %
age	25	25	3	100 %
children	4	4	0	100 %
status	0	0	0	100 %
gender	1	1	0	100 %

Se encontraron un total de 48 tuplas en 11 columnas con valores anómalos (algunas de las tuplas pueden duplicarse por poseer más de una columna anómala). Solamente la columna edad contiene valores nulos que no fueron considerados en el análisis.

El segundo paso del procedimiento consistió en la aplicación del algoritmo de vecinos más cercano KNN sobre cada columna del perfil del cliente. Esto fue útil para obtener una calificación de la anomalía con respecto al conjunto de datos. Los resultados de la aplicación del algoritmo establecieron que el 80 % de las columnas contienen valores anómalos, y un restante 20 % resultaron una mezcla entre valores anómalos y falsos-positivos. Por ejemplo, en el caso de la columna "age" 1154 son tuplas de personas de hasta 67 años, 3 valores son nulos y 23 valores corresponden a personas con más de 102 años. La identificación previa por tablas de frecuencia, separó los valores nulos y los valores de personas con más de 102 años, sin embargo, luego de aplicar el algoritmo KNN, se obtuvieron personas con menos de 102 años, como casos de valores anómalos. Esto determinó la inclusión de una siguiente fase que separa falsos-positivos de aquellos realmente anómalos.

Similar al artículo de los autores H Kuna, Rambo y Caballero (2012), el algoritmo que se aplicó fue el K-means para separar los valores anómalos de los falsos positivos. El método consistió en la aplicación del algoritmo con un parámetro k=2, que considera 2 clústeres: grupo de datos falsos-positivos, grupo de valores anómalos. La diferencia con el trabajo de Kuna radica en la inclusión de una fórmula que mide las distancias de cada clúster al punto establecido por el promedio de distancias totales. A través del indicador del grado de distancia, se identificaron el clúster anómalo y el clúster de falsos-positivos. Luego de afinar el parámetro del grado de distancia, se logró una efectividad del 100 % en la clasificación de valores atípicos y de falsos-positivos.

La calificación del grado de relación será importante para el usuario final ya que puede relacionarla con el valor anómalo y así establecer un grado de influencia en el resultado final.

5. Conclusiones y recomendaciones

Este artículo trata sobre la generación de un método híbrido para descubrir valores atípicos considerando el uso de una prueba estadística y técnicas de minería de datos. Como parte de la validación del método, se analizó una base de datos que otorga créditos financieros, la cual estuvo compuesta por atributos que describen el perfil del cliente con relación al monto otorgado de crédito. El proceso de análisis de datos de cada uno de los atributos identificó aquellas columnas que tienen relevancia con respecto a la columna de salida, siendo esta característica importante para la calificación del grado de la anomalía, a través de la técnica de vecinos más cercanos K-NN. Los valores mínimos de relevancia y de anomalía se pueden configurar según sea el caso de estudio, lo que permite analizar varios estratos de salida. Será necesario entonces encontrar los valores óptimos de estos parámetros según sea el caso de estudio. El procedimiento propuesto fue adecuado para la identificación de valores atípicos en un conjunto de datos con diversidad en la variabilidad de los datos de tipo nominal. La fortaleza del proceso radica en que se combinan varias técnicas que identifican con gran precisión los valores atípicos sin importar la presencia de valores nulos. Es decir, la metodología tiene dos aspectos importantes que lo diferencian de otros métodos: una prueba estadística que otorga relevancia a las columnas y dos técnicas de minería de datos que separan las anomalías de los datos normales. Una vez que los parámetros y los filtros están adecuadamente configurados en cada parte del método, en el caso particular se llegó a detectar el 100 % de los valores atípicos. Como trabajo futuro, se recomienda además de encontrar anomalías a nivel de columna, combinar columnas para determinar patrones relacionados con la columna de salida. También es necesario ejecu-

tar evaluaciones adicionales con otros conjuntos de datos en ámbitos de diversa índole para validar la efectividad del método y la eficiencia en el uso de los recursos computacionales. Se recomienda también un estudio de las técnicas alternativas para mejorar la configuración de la variable *k* del algoritmo k-NN de forma automatizada.

Bibliografía

Aldahdooh, R. T. y Ashour, W. M. (2013). DIMK-means Distance-based Initialization Method for K-means Clustering Algorithm. Intelligent Systems and Applications, 5 (2): 41-51.

Amer, M. y Goldstein, M. (2012). Nearest-Neighbor and Clustering based Anomaly Detection Algorithms for RapidMiner. *Proceedings of the 3rd RapidMiner Community Meeting and Conferernce (RCOMM 2012)*, 1-12.

Arce, D., Lima, F., Orellana, M., Ortega, J. y Sellers, C. (2018). Discovering behavioral patterns among air pollutants : A data mining approach (Descubriendo patrones de comportamiento entre contaminantes del aire : Un enfoque de minería de datos). *Enfoque UTE* 9 (4): 168-179.

Atkinson, A. C. (1981). Identification of Outliers. *Biometrics*, 37 (4): 860-861.

Bansal, R., Gaur, N. y Singh, S. N. (2016). Outlier Detection: Applications and techniques in Data Mining. *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)*, 373-377. https://doi.org/10.1109/CONFLUENCE.2016.7508146

Bhattacharyya, S., Jha, S., Tharakunnel, K. y Westland, J. C. (2011). Data mining for credit card fraud: A comparative study. *Decision Support Systems*, 50 (3): 602-613. https://doi.org/10.1016/j.dss.2010.08.008

Dang, T. T., Ngan, H. Y. T. y Liu, W. (2015). Distance-based k-nearest neighbors outlier detection method in large-scale traffic data. *International Conference on Digital Signal Processing, DSP*, *2015-September*, 507-510. https://doi.org/10.1109/ICDSP.2015.7251924

Ganji, V. R. (2012). Credit card fraud detection using anti-k nearest neighbor algorithm. *International Journal on Computer Science and Engineering*, 4 (6): 1035-1039.

Gol, M. y Abur, A. (2015). A modified Chi-Squares test for improved bad data detection. *2015 IEEE Eindhoven PowerTech, PowerTech 2015*, (1): 1-5. https://doi.org/10.1109/PTC.2015.7232283

Gu, Y., Ganesan, R. K., Bischke, B., Bernardi, A., Maier, A., Warkentin, H., … Dengel, A. (2017). Grid-based outlier detection in large data sets for combine harvesters. *Proceedings-2017 IEEE 15th International Conference on Industrial Informatics, INDIN 2017*: 811-818. https://doi.org/10.1109/INDIN.2017.8104877

Hassanat, A. B., Abbadi, M. A. y Alhasanat, A. A. (2014). Solving the Problem of the K Parameter in the KNN Classifier Using an Ensemble Learning Approach. *International Journal of Computer Science and Information Security (IJCSIS)*, 12 (8): 33-39. https://doi.org/10.1007/s00500-005-0503-y

Hassanat, A. B., Abbadi, M. A., Altarawneh, G. A., y Alhasanat, A. A. (2015). A SURVEY OF OUTLIER DETECTION IN DATA MINING. *International Journal of Advance Engineering and Research Development*, 3 (01). https://doi.org/10.21090/ijaerd.ncrretcs06

Khan, M. A., Pradhan, S. K. y Fatima, H. (2017). Applying Data Mining Techniques in Cyber Crimes. *2nd International Conference on Anti-Cyber Crimes*, 2-5. https://doi.org/doi: 10.1109/Anti-Cybercrime.2017.7905293

Kuna, H, Rambo, A. y Caballero, S. (2012). Procedimientos para la identificación de datos anómalos en bases de datos. *Proceedings Of*. Retrieved from http://sistemas.unla.edu.ar/sistemas/gisi/papers-HK/procedimientos para la identidficacion de datso anomalos en bases de datos.pdf

Kuna, Horacio, Pautsch, G., Rambo, A., Rey, M., Cortes, J., Rolón, S. y Informática, D. De. (2013). Procedimiento de Explotación de Información para la Identificación de Campos anómalos en Base de

Datos Alfanuméricas. *Revista Latinoamericana de Ingeniería de Software*, 1 (3): 102-106. Retrieved from http://sistemas.unla.edu.ar/sistemas/redisla/ReLAIS/relais-v1-n3-p-102-106.pdf

Malini, N. y Pushpa, M. (2017). Analysis on credit card fraud identification techniques based on KNN and outlier detection. *Proceedings of the 3rd IEEE International Conference on Advances in Electrical and Electronics, Information, Communication and Bio-Informatics, AEEICB 2017*: 255-258. https://doi.org/10.1109/AEEICB.2017.7972424

Mandhare, H. y Idate, S. (2017). A Comparative Study of Cluster Based Outlier Detection, Distance Based Outlier Detection and Density Based Outlier Detection Techniques. *International Conference on Intelligent Computing and Control Systems*: 931-935.

Monamo, P. M., Marivate, V. y Twala, B. (2017). A multifaceted approach to Bitcoin fraud detection: Global and local outliers. *Proceedings - 2016 15th IEEE International Conference on Machine Learning and Applications, ICMLA 2016*, 188–194. https://doi.org/10.1109/ICMLA.2016.19

Onan, A. (2017). A K-medoids based clustering scheme with an application to document clustering. *2nd International Conference on Computer Science and Engineering, UBMK 2017*: 354-359. https://doi.org/10.1109/UBMK.2017.8093409

Ougiaroglou, S., Evangelidis, G. y Dervos, D. A. (2014). FHC: An adaptive fast hybrid method for k-NN classification. *Logic Journal of the IGPL*, 23 (3): 431–450. https://doi.org/10.1093/jigpal/jzv015

Rosero-Montalvo, P. D., Umaquinga-Criollo, A. C., Flores, S., Suarez, L., Pijal, J., Ponce-Guevara, K. L., … Moncayo, K. (2018). Neighborhood criterion analysis for prototype selection applied in WSN data. *Proceedings-2017 International Conference on Information Systems and Computer Science, INCISCOS 2017*, *2017-Novem*: 128-132. https://doi.org/10.1109/INCISCOS.2017.47

Sinwar, D. y Dhaka, V. S. (2015). Outlier detection from multidimensional space using multilayer perceptron, RBF networks and pattern clustering techniques. *Conference Proceeding-2015 International Conference on Advances in Computer Engineering and Applications, ICACEA 2015*: 573-579. https://doi.org/10.1109/ICACEA.2015.7164757

Sumaiya Thaseen, I. y Aswani Kumar, C. (2017). Intrusion detection model using fusion of chi-square feature selection and multi class SVM. *Journal of King Saud University - Computer and Information Sciences*, 29 (4): 462-472. https://doi.org/10.1016/j.jksuci.2015.12.004

Yan, K., You, X., Ji, X., Yin, G. y Yang, F. (2016). A hybrid outlier detection method for health care big data. *Proceedings - 2016 IEEE International Conferences on Big Data and Cloud Computing, BDCloud 2016, Social Computing and Networking, SocialCom 2016 and Sustainable Computing and Communications, SustainCom 2016*: 157-162. https://doi.org/10.1109/BDCloud-SocialCom-SustainCom.2016.34

Zhang, H. y Wang, L. (2018). An information-Theoretic outlier detection method for prescription data. *2017 3rd IEEE International Conference on Computer and Communications, ICCC 2017*, *2018-January*: 2361-2365. https://doi.org/10.1109/CompComm.2017.8322957